Please carefully read and follow the general instructions regarding computing assignments. Failing to meet the requirements might lead to penalties. [https://moodle.uef.fi/mod/page/view.php?id=2258688](https://moodle.uef.fi/mod/page/view.php?id=2258688)

If you suspect that something is wrong with some task instructions, please contact the lecturer.

If you face persistent issues while working on a computing assignment, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

Attached material.

creditDE.csv  Credit, multi-dimensional binary classification dataset
classification_resources.py  some potentially useful code snippets

The dataset is a variant Statlog dataset from the UCI Machine Learning Repository.

The dataset can be loaded using the load_csv function provided in classification_resources.py:

```python
import classification_resources

dataset, h, c = load_csv("creditDE.csv")
```

Imports of external libraries other than those that appear in the classification_resources.py file are not allowed.

**Task 1.** Implement the RBF kernel for the SVM algorithm. Divide the Credit dataset into training and test subsets in proportions 3/4–1/4 at random, i.e. assign one fourth of instances, chosen at random, to the test dataset and the rest to the training dataset. Plot the ROC curve and compute the AUC for the linear SVM and for the SVM with RBF kernel (setting $\sigma = 1$, for example).

**Task 2.** Run a comparative evaluation of the linear SVM vs. SVM with RBF kernel on the Credit dataset, first with cross-validation, then with bootstrap samples.

Run 10 rounds of cross-validation with 5 folds on the Credit dataset, training a linear SVM classifier on one hand, and a SVM with RBF kernel, using the same data splits in both cases. Estimate the mean and variance of the classifiers’ accuracy across the successive rounds and determine which classifier performs best and whether the performance gap is statistically significant.

**Task 3.** Repeat, using 50 bootstrapped samples instead of cross-validation. In each round, train both classifiers on a bootstrapped sample and test them on the original data.

Compare your results to those obtained in the previous task and discuss.

---

[https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)](https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))